

結合音節文字の符号化方式の分類

三 上 喜 貴*

Classification of Coding Models for
Combining Syllabics

Yoshiki MIKAMI

Key Words: character code, combining character, syllabics

目 次

- 1 本稿の目的
- 2 文字集合
 - 2.1 文字集合の特徴と構成要素
 - 2.2 子音文字 (Consonant syllabics)
 - 2.3 母音文字 (Vowel characters) と母音記号 (Vowel signs)
 - 2.4 脱母音記号 (Devowelizer)
 - 2.5 子音記号 (Consonant Signs)
 - 2.6 補助記号 (声調記号等) (Diacritics)
 - 2.7 数字 (Numerals)
 - 2.8 文節記号 (Punctuation Marks)
 - 2.9 その他の非表音記号 (Miscellaneous Non-phonetic Symbols)
- 3 結合規則
 - 3.1 音節の基本構成
 - 3.2 開音節の表現
 - 3.3 多重子音の表現
 - (1) 脱母音記号
 - (2) 子音記号

原稿受付：平成11年5月21日

*長岡技術科学大学 計画・経営係

(3) 二階建て文字

(4) 合字

4 結合音節文字の符号化方式

4.1 符号化方式の類型

4.2 GCGO: Graphical Coding Graphical Order

4.3 GCPO: Graphical Coding Pronunciation Order

4.4 GVPC: Graphic Vowel-coding Phonetic Consonant-coding

4.5 PC: Phonetic Coding

4.6 SC: Syllabic Coding

5 まとめ

1 本稿の目的

言語学の分野では、一般に文字の類型について表1のような分類が行われる²。アジア地域にはこれらの各類型に属する全ての種類の文字体系が存在する。東アジア各国で使用されている漢字は表語文字（logogram）であり、韓国語のハングルや日本語のカナは音節文字（syllabics）である。他方、南アジア及び東南アジア地域では、植民地時代を通じて国語表記のローマ字化が進行した幾つかの国で単音文字（alphabet）であるラテン・アルファベットが使用されているものの、大部分の地域では古代インド文字を共通の起源として発達した音節文字が使用されている。

表1 文字体系の分類

類型		文字集合	文字の例
表音文字 phonogram	単音文字 alphabet	ラテンアルファベット	Bahasa Indonesia
	音節文字 syllabics	ハングル文字*	
		ひらがな*	にほんご
		カタカナ*	イングリッシュ
		デーヴァナーガリー文字	भारत
		クメール文字	ខ្មែរ
		ミャンマー文字	မြန်မာ
表語文字 logogram	或いは表意文字 ideograph	漢字	普通話、国語

ここで「音節文字」とは「音節を単位として表記される表音文字」と定義されるが、多くの場合、複数の図形要素が一定の規則に従って結合し、一つの音節を表す。ハングル文字やインド系文字は共にこのような性質を有している。本稿では、音節文字のうちこのような性質を有するこれらの文字を結合音節文字 (combining syllabics) と呼ぶこととする。

結合音節文字の符号化方式には、個々の図形要素を単位として符号化するか或いは結合した後の音節を単位として符号化するか等の選択肢があり、更に前者の方法によった場合、符号化された図形要素をどのような順序に従って配列するのかという選択肢も派生する。アジア各国ではこれまでに幾つかの国家的な文字符号化規格が生み出されてきたが、本稿でみるように、これらはそれぞれに異なる符号化原理に基づいて設計されている。表語文字や単音文字の場合にはこのような選択の余地はなく、文字とその符号化表現とが一對一に対応しているのと対照的である。

筆者は、1997年以来、工業技術院標準部の支援を受けて実施されている「多言語情報処理技術開発プロジェクト」の一環として、アジア関係国の専門家との国際共同作業による文字符号化方式に関する共通ガイドラインの策定に取り組んでいる³。これらの地域におけるコンピュータの利用状況は先進国と比べれば未だ限られたものであるが、今後、自国語で必要とする文字を単独で或いは多言語環境の下で使用する必要性が高まっていくことは明らかであり、近い将来に予想されるこうした本格的な多言語情報処理への要請に応えるためには、既存の符号化方式の相違点や共通点を明確にし、処理系が満足すべき条件を明らかにすることが必要である。また、異なる符号化方式を多面的に評価し、望ましい符号化方式に関するガイドラインを策定することは、これから新たに文字符号化標準の策定に取り組もうとする多くの国にとって貴重な技術的支援を提供することとなる。

現在のところ、UNICODE⁴やISO/IEC 10646⁵の開発作業をきっかけとして多くの国がこの問題に対する関心を高めており、各国の専門家が符号化方式についての研究・検討を行っているが⁶、これを横断的な視点から取り扱う試みはまだ行われていない。

こうした観点に立ち、共通ガイドライン策定のための第一段階として、本稿は、主としてインド系文字を対象として、結合音節文字の符号化方式に関する分類を試みようとするものである。

なお、表2にアジアにおけるコンピュータ及びインターネットの普及状況を示し、表3にアジア各国で使用されている公用言語とその表記に使用する文字、文字符号化規格の有無を示す。両表から明らかとなっており、結合音節文字の使用地域のうち、国家的なレベルで文字符号化規格が制定されているのはインド、タイ、スリランカ、バングラデシュ及び韓国のみであり、これら以外の地域では複数の、互換性のない文字符号表が混在して使用されている状況にある。また国家的なレベルでの公用語となっていない、いわゆる少数民族語の使用文字を含めれば、結合音節文字の符号化に関する標準化はその端緒についたばかりの状況にあるといえる。

表2 アジア各国・地域のコンピュータ化の現状

国・地域	(A) 人口(千人) (1997)	(B) Internetドメイン数 (1998.7)	(C) = (B) / (A) 人口千人当り Internetドメイン数
(東アジア)			
日本	125,761	1,352,200	10.75
中国	1,232,083	494	0.00
韓国	45,545	174,800	3.84
北朝鮮	22,466	NA	NA
モンゴル	2,354	17	0.01
(東南アジア)			
ブルネイ	300	740	2.47
カンボジア	10,702	58	0.01
インドネシア	196,813	10,691	0.05
ラオス	5,035	0	0.00
マレーシア	20,565	40,758	1.98
ミャンマー	45,922	0	0.00
フィリピン	71,899	7,602	0.11
シンガポール	3,044	59,469	19.54
タイ	60,003	25,459	0.42
ベトナム	75,181	25	0.00
(南アジア)			
バングラデシュ	120,073	NA	NA
ブータン	1,812	2	0.00
インド	939,409	10,436	0.01
モンディブ	263	70	0.27
ネパール	21,127	123	0.01
パキスタン	134,146	1,923	0.01
スリランカ	18,300	580	0.03
結合音節文字使用国			
(太字)の合計	1,211,681	36,658	0.03
アジア合計	3,152,803	1,685,447	0.53

(出典) 人口は「日本国勢図会98/99年版」によった。インターネット・ドメイン数はNetwork Wizards調査による。URL <http://www.nw.com>

結合音節文字の符号化方式の分類

表3 アジア地域における言語と文字

国・地域	主たる言語	左記に用いられる 文字集合	文字コード 規格の有無
(東アジア)			
日本	日本語	カナ 漢字	JIS C6220-1969 JIS C6226-1978等
中国	中国語	漢字(簡体字)	GB2321-1980等
韓国	韓国語	ハングル, 漢字等	KS C5601-1987
北朝鮮	朝鮮語	ハングル, 漢字等	KPS 9566: 1997
モンゴル	モンゴル語	キリル文字(※)	
(東南アジア)			
ブルネイ	マレー語	ラテン文字	—
	英語	ラテン文字	—
カンボジア	カンボジア語	クメール文字	制定の動き
インドネシア	インドネシア語	ラテン文字	—
ラオス	ラオ語	ラオ文字	制定の動き
マレーシア	マレー語	ラテン文字	—
ミャンマー	ミャンマー語	ミャンマー文字	98年に制定
フィリピン	ピリピノ語	ラテン文字(※)	—
シンガポール	英語	ラテン文字	
	中国語	漢字	
	マレー語	ラテン文字	
	タミル語	タミル文字	
タイ	タイ語	タイ文字	TIS 620:1986
ベトナム	ベトナム語	ラテン文字(※)	TCVN 5412-1993
(南アジア)			
バングラデシュ	ベンガル語	ベンガル文字(※)	BDS 1520: 1995
ブータン	ゾンカ語	ゾンカ文字	不明
インド	ヒンディー語, コンカニ語, ネパール語, マラティ語, サンスクリット語	デーヴァナーガリー文字	IS 13194:1991 (ISCI1)
	パンジャブ語	パンジャブ文字	
	グジャラティ語	グジャラティ文字	
	オリヤ語	オリヤ文字	
	ベンガル語	ベンガリ文字	
	アッサム語	アッサム文字	
	テルグ語	テルグ文字	
	カンナダ語	カンナダ文字	
	マラヤラム語	マラヤラム文字	
	タミル語	タミル文字	
モンルディブ			不明
ネパール	ネパール語	ネパール文字(※)	制定の動きあり
パキスタン	ウルドゥー語	ウルドゥー文字	不明
スリランカ	シンハリ語	シンハラ文字	SLS 1134:1996
	英語	ラテン文字	
	タミル語	タミル文字	

(注) ※印は若干の追加文字を必要とすることを意味する。文字コード規格については、利用可能となった時期を示すために、最初の版の発行年を付して示した。これらの多くはその後改訂を加えられている。

(出典) 「主たる言語」については外務省編「世界の国々」1998年版によった。

2 文字集合

2.1 文字体系の特徴と構成要素

南アジア及び東南アジアで使用されている非ラテン文字のほとんどは古代インドの文字を起源として発展してきた文字であり、これらをその歴史的起源に着目して「インド系文字」と呼ぶ⁷⁾。インド系文字は通常音節文字に分類され、文字が単独で音節を表すことができるが、日本語のカナのような完全音節文字と異なり、全ての音節と1対1に対応付けるに足る数の独立した文字が揃っているわけではなく、基本となる文字の周りに幾つかの図形要素を結合させて様々な音節を表現する。

ISO/IEC 10646で定義された用語⁸⁾に従えば、このことは「基底文字の周囲に結合文字を組み合わせて一つの音節を表す合成列が形成される」と表現される。

結合音節文字の文字集合の要素であって基底文字とみなされるのは子音文字であり、幾つかの言語では、独立して表記される場合の母音文字がこれに加わる。表音的な機能を持つその他の図形要素は総て結合文字である。これらの中には、多重子音を作るために子音文字と結合して用いられる子音記号、母音音価を変更する機能を持つ母音記号、声調記号等が含まれる。ISOの定義に従えばこれらの結合文字も「文字」であるが、本稿ではこれらの子音記号、母音記号、補助記号等、「記号」と呼んで区別する。文節記号や数字等の非表音的な機能を持つ図形要素も結合して用いられることはないからISOの定義に従えば基底文字となるが、本稿ではこれらも「文字」とは呼ばず「記号」と呼ぶこととする。従って本稿では、結合音節文字における文字集合の構成単位を次のように分類する。

表4 結合音節文字の文字集合における構成単位

結合/非結合	構成単位	略号	機能
基底文字 =非結合文字	子音文字(consonant syllabics)	C	子音を含む音節を表す
	母音文字(vowel characters)	V	単独で母音を表す
	数字(numerals)	Sn	数字を表す
	文節記号(punctuation marks)	Sp	文の区切りを表す
	その他の非表音記 (miscellaneous non-phonetic symbols)	Sm	繰り返し、省略等を表す
結合文字	子音記号(consonant signs)	C	子音文字と結合して多重子音を作る
	母音記号(vowel signs)	V	子音文字と結合して母音を変化させる
	脱母音記号(devowelizer)	X	子音文字と結合して母音を消去する
	補助記号(diacritics)	D	音節全体の高低・長短等を変化させる

以下において、上記の文字、記号の機能を説明し、次いでその結合規則を説明する。

2.2 子音文字 (Consonant syllabics)

インド系文字を使用している言語の系統はインド・アーリア語族、ドラビダ語族、モン・クメール語族、シナ・タイ語族、オーストロネシア語族等、多岐にわたり、各言語で使用される子音の種類は元来異なる。しかしインド文化との交流過程でサンスクリット・パーリ語からこれらの言語へと多くの語彙が輸入され、同時にサンスクリット・パーリ語の子音も導入された。このため現在使用されている子音文字は概ねサンスクリット・パーリ語の子音に対応する30～40種類の子音文字で構成されている。インド系主要言語の子音文字の一覧表を表5に示した。

結合音節文字の体系においては、子音文字が特定の既定値（多くのインド系文字において「a」音である）に固定された固有の母音音価（inherent vowel）を伴うところに特徴がある⁹。従って子音文字とはいえ、その音価は子音と母音の両方を含み、単独で音節を表すことのできる音節文字であるから「子音音節文字」（Consonant Syllabics）と呼ぶのが相応しいが、本稿では簡略のため子音文字と呼ぶ。

結合音節文字における基本文字集合はこの「母音の音価が特定の既定値に固定された子音文字」と後述する「母音文字」とから構成されている。日本語のカナに喩えれば「か、さ、た、な、は、ま、や、ら、わ」と「あ、い、う、え、お」だけから構成されている文字体系と考えれば良い。

2.3 母音文字 (Vowel characters) と母音記号 (Vowel signs)

インド系文字では子音文字に含まれる母音が既定の音価に固定されているために、既定値と異なる母音へ変更する場合には母音の既定値を変更するための特別の記号を必要とする。これが母音記号（Vowel Signs）である。子音文字に付加されてその母音音価を変更するものであることから子音変更記号（Consonant modifiers）と呼ばれることもある。

一方、母音が単独で語頭或いは音節頭に現れる場合、母音文字が使用される。インド系文字の中には母音文字を持たないものもある。表6に主なインド系文字について母音文字及び母音記号の一覧を示した。

2.4 脱母音記号 (Devowelizer)

インド系文字では子音文字が既定の音価の母音を含んでいるために、

- ・ 多重子音を作る際の先行子音からの母音の削除
- ・ 終止音となる子音からの母音の削除

等の場合に母音を削除するための特別の記法が発達した。これが脱母音記号 (Devowelizer or Vowel Omission Sign) である。或いはパーリ語の呼び名でViramaと呼ばれることもある。脱母音記号はほとんどのインド系文字に存在するが、脱母音記号を表記しない文字体系もある。

2.5 子音記号 (Consonant Signs)

結合子音を作る子音文字については、単独で音節を表す時の図形とは別の記号がある場合がある。これを子音文字と区別して「子音記号」(Consonant signs) と呼ぶ。

2.6 補助記号 (声調記号等) (Diacritical Marks)

子音・母音を表す文字・記号で表現される基本的な音を強弱、高低、長短等の面から変化させることを表す記号として様々な補助記号 (Diacritical Marks) がある。声調¹⁰を示す声調記号 (Tone Marks) もその一つである。本稿で取り上げる言語のうち、声調言語 (tone language or tonal language) といわれるのはタイ語、ベトナム語、クメール語などであり、これらの言語では声調を区別するための記号が明示的に文字列中に書き込まれる。声調記号も補助記号の一つであるから、ここでは声調記号とその他の補助記号とを区別しない。補助記号は母音を修飾するものであることから母音変更記号 (Vowel modifiers) と呼ばれることもある。

2.7 数字 (Numerals)

全てのインド系文字はアラビア数字と異なる独自の数字を持っている。基本的には十進数の体系であるが、位取りや桁の表示方法などに関してアラビア数字とは異なる表記法を採用するものもある。

2.8 文節記号 (Punctuation Mark)

ほとんど全てのインド系文字に独自の文節記号 (Punctuation Marks) がある。但し次第にラテンアルファベットの文節記号 (カンマとピリオド) を用いて表記する例も増えている。

2.9 その他の非表音記号 (Miscellaneous Non-phonetic Symbols)

以上の他、インド系の文字集合には「繰り返し」、「省略」などを表す非表音的な機能を持つ記号類が含まれる。

表5 主要なインド系文字の子音文字一覧

	1	2	3	4	5	6	7	8	9	10
IPA	k	k ^h	g	g ^h	ŋ	c	c ^h	j	j ^h	ɳ
Devanagari	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
Khmer	ក	ខ	ក	ឃ	ង	ច	ឆ	ជ	ឈ	ញ
Lao	ກ	ຂ		ຄ	ງ	ຈ			ຊ	ຍ
Myanmar	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ည	ဉ
Sinhala	ක	ඛ	ග	ඝ	ඞ ඪ	ච	ඡ	ජ	ඣ	ඤ ඳ ඵ
Tamil	க				ங	ச		ஜ		ஞ
Thai	ก	ข ช	ค ฌ	ฌ	ง	จ	ฉ	ช	ฌ	ญ

	11	12	13	14	15	16	17	18	19	20
IPA	t	t ^h	d	d ^h	ɳ	ʈ	ʈ ^h	ɖ	ɖ ^h	n
Devanagari	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
Khmer	ត	ថ	ដ	ឆ	ណ	ត	ថ	ទ	ជ	ន
Lao	ຕ	ຖ	ດ	ຖ	ນ					
Myanmar	တ	ထ	ဒ	ဉ	ဏ	တ	ထ	ဒ	ဓ	န
Sinhala	ට	ඨ	ඩ	ඪ	ණ ඬ	ත	ථ	ද	ධ	න ඳ
Tamil	ட				ண	த				ந ன
Thai	ด ต	ฐ	ท	ฒ	ณ	ด ต	ถ	ท	ธ	น

表5 主要なインド系文字の子音文字一覧(続き)

	21	22	23	24	25	26	27	28	29	30
IPA	p	pʰ	b	bʰ	m	y	r	l	l̥	v
Devanagari	प	फ	ब	भ	म	य	र	ल	ळ	व
Khmer	ប	ផ	ព	ភ	ម	យ	រ	ល		វ
Lao	ປ	ຝ		ພ	ມ	ຢ	ຣ	ລ		ວ
	ປ	ຝ		ພ						
Myanmar	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ		ဝ
Sinhala	ප	ඵ	බ	භ	ම	ය	ර	ල	ළ	ව
					ඔ					
Tamil	ப				ம	ய	ர	ல	ள	வ
							ற		ழ	
Thai	ป	ผ	พ	ภ	ม	ย	ร	ล		ว
	ป	ผ	พ				(ร)	(ล)		

	31	32	33	34	35	36	37	38	子音文字数
IPA	s	ʃ	s	h	la	a/o	h	f	
Devanagari	श	ष	स	ह					34
Khmer			ស	ហ	ឡ	អ			33
Lao			ສ	ຫ		ອ	ຮ		27
Myanmar			သ	ဟ	ဠ	အ			33
Sinhala	ශ	ඡ	ස	හ				ඟ	41
Tamil		ஷ	ஸ	ஹ					22
Thai	ศ	ษ	ส	ห	ฬ	อ	ฮ		44

(注) 音価は各子音文字に対応する子音音価を国際音声字母 (IPA) で示したものである。サンスクリット／パリー語の子音順序に合わせて配列したため空白の欄もある。逆に一つの欄に複数の文字があるのは近似音の文字を同じ欄内で示したためである。タイ文字の /r/, /l/ 音を表す文字のうち下段に示した文字 (括弧書きした文字) は通常母音として扱われている。

表6 インド系文字の母音文字と母音記号

(A) Devanagari

音	a	ā	i	ī	u	ū	ri	ē	ai	ē	ō	au	ō
母音文字	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ँ	ओ	औ	ऑ
母音記号	क	का	कि	की	कु	कू	कृ	के	कै	कँ	को	कौ	कॉ

音	rī	li	lī										
母音文字	ऌ	ल	लृ										
母音記号	कृ	कृ	कृ										

(B) Khmer

akosa音	a	ā	e	ei	oe	oe	o	ō	uo	aoe
kosa音	o	ea	i	ī	oe	oe	u	ū	uo	oe
母音記号	ក	កា	កី	ក្រ	ក្រ	ក្រ	កុ	កូ	កួ	កើ
母音文字			ឥ	ឿ			ឧ			
akosa音	ua	ie	ē	ae	ai	ao	aow	om	om	am
kosa音	ua	ie	ē	ae	ay	ō	eow	um	om	eam
母音記号	ក្រើ	ក្រៀ	កើ	ក្រៃ	ក្រៃ	ក្រៅ	ក្រៅ	កុំ	កំ	កាំ
母音文字				ឯ	ឮ	ឱ	ឱ			
akosa音	ah	oh	eh	aoh	roe	roē	loe	loē		
kosa音	eah	uh	eh	oh						
母音記号	កៈ	កៈ	កៈ	កៈ						
母音文字					ឬ	ឬ	ព	ព		

(C) Lao

音	a	ā	i	ī	u	ū	e	ē	ō	i
母音記号	ກະ	ກາ	ກິ	ກີ	ກຸ	ກູ	ກຸ	ກູ	ກະ	ກາ
音										
母音記号	ແກະ	ແກ	ໂກະ	ໂກ	ເກະ	ກໍ	ເກີ	ເກ້	ເກ້ະ	ເກ້
音										
母音記号	ກ້ວະ	ກັວ	ເກືອ	ເກືອ	ໄກ	ໄກ	ກໍ່	ເກົາ		

(D) Myanmar

音	a	ā	i	ī	u	ū	e	ē	ō	au		
母音文字	အ	အာ	အိ	အီ	အု	အူ	အေ	အဲ	အော	အော်	အံ	အို
代替形			အူ	အို	အု	အူ	အေ		အူ	အော်		အာ:
母音記号	က	ကာ	ကိ	ကီ	ကု	ကူ	ကေ	ကဲ	ကော	ကော်	ကံ	ကား

(E) Sinahala

音	a	ā	æ	ǣ	i	ī	u	ū	ɪ
母音文字	අ	ආ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	සා
母音記号	-	-ා	-ැ	-ෑ	-ි	-ී	-ු	-ූ	-a
音	ī	ilu	ilū	e	ē	ai	o	ō	au
母音文字	සා	ඬ	ඬා	එ	ඒ	ඒ	ඔ	ඔ	ඔා
母音記号	-aa			-ෙ	-ේ	-ෙඒ	-ෙඔ	-ෙඔ	-ෙඔා

(注) æとǣに対応する二つの母音は、インド系文字を使用する他の言語には見られないシンハラ語固有の母音。
ilu、ilūに対応する母音は現代のシンハラ語では使用されていない。

(F) Tamil

音	a	ā	i	ī	u	ū	e	ē
母音文字	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ
母音記号	-	-ஈ	-ி	-ீ	-ு	-ூ	-ெ	-ே
音	ai	o	ō	au	u			
母音文字	ஐ	ஔ	ஓ	ஔ	உ			
母音記号	ஐ- உ	ஔ- ஈ	ஓ- ஈ	ஔ- ஈ	- ு			

(G) Thai

音	a	ā	i	ī	u?	uu	u?	uu
母音文字	กะ	กา	กิ	กี	กี	กี	กุ	กู
音	e?	ee	e?	ee	o?	oo	o?	oo
母音文字	เกาะ	เก	เกาะ	เก	โกะ	โก	เกาะ	กอ
音	o?	oo	ia?	ia	ua?	ua	ua?	ua
母音文字	เกอะ	เกอ	เกียะ	เกีย	เกือะ	เกือ	กัวะ	กัว
音	ru?	ruu	am	ai	ai	ao		
母音文字	ฤ	ฤา	กำ	ไก	ไก	เกา		

3 結合規則

3.1 音節の基本構造

音節はその中心となる母音を必ず含み¹¹、これに子音が組み合わされて構成される。音節の先頭に位置する子音を頭子音と呼び、音節末尾に位置する子音を末子音と呼ぶ。末子音を持つ音節を閉音節 (closed syllable) と呼び、末子音を持たずに母音で終わる音節を開音節 (open syllable) と呼ぶ。音節中の子音は単独で或いは複数の子音の組み合わせとして現れる。子音の組み合わせられたものを多重子音 (consonant conjunct) と呼ぶ。音節構文規則の一般形を拡張されたBNF形式を用いて書けば式 1～7 となる。

<Word>	::= {<Syllable>}	式 1
<Syllable>	::= <Open-Syllable> <Closed-Syllable>	式 2
<Open-Syllable>	::= <Vowel-Syllable> <Cons-Vowel-Syllable>	式 3
<Closed-Syllable>	::= <Open-Syllable> <Cons-Conjunct>	式 4
<Vowel-Syllable>	::= <Vowel> [<Diacritic>]	式 5
<Cons-Vowel-Syllable>	::= <Cons-Conjunct> <Vowel> [<Diacritic>]	式 6
<Cons-Conjunct>	::= {<Consonant>}	式 7

{ }	1 回以上の繰り返しがあることを示す
[]	場合により付加されることを示す

このようにして構成される一つの音節を一つの文字で表す文字が音節文字であり、結合音節文字では複数の文字・記号を組み合わせで一つの音節を表す。韓国のハングル文字は頭子音 (初声), 母音 (中声), 末子音 (終声) を表す字母 (Jamo) の組み合わせで表現される音節文字であるから結合音節文字である。これに対して、日本語のカナは文字を音節要素へと分解することができないから非結合の音節文字といえる¹²。

文字, 記号から音節を構成するに際しては一定の構文規則 (syntax) がある。ここで構文規則には二つの側面があることに注意する必要がある。一つは論理的な構文を表すものであり、もう一つは空間的配置における結合規則である。アルファベットの場合には連続する文字が直線的, 一次元的に配置されているために二つの構文規則を明確に区別する必要はないのに対して¹³, 結合音節文字においては子音文字を中心としてその周囲に子音記号, 母音記号, 声調記号等が二次元的に結合して配置されているため, 二つの構文規則を分けて考

える必要がある。この点に留意しながら具体例を示しつつ以下に結合規則を概説する。

3.2 開音節の表現

【論理的構成】

まず音節表現の基本型として開音節を示す。結合音節文字においては、母音文字(V)、子音文字(C)はともに単独で音節を表すことができる。しかし子音文字は母音が既定値に固定されているから、既定値以外の音に変更する場合には子音文字に母音記号(V)を付加してこれを示す。また母音は補助記号(D)によって修飾される場合がある。補助記号はそれ自体で何らかの音価を持つものではなく、音節全体の音を修飾するものであるから論理式の上では最後に置く。

先の音節式5と6を子音文字、母音文字、母音記号、補助記号から構成される構文規則として書き換えると次のようになる。なお、表記にあたっては一文字の略記号を用いた簡潔な表記を行う。

<Vowel-Syllable>	::= V[D]	式 5 a
<Cons-Vowel-Syllable>	::= C[<u>V</u>][D]	式 6 a
V	: 母音文字 (vowels)	
<u>V</u>	: 母音記号 (vowel signs)	
C	: 子音文字 (consonants)	
D	: 補助記号 (diacritics)	

またインド系文字の中には、母音文字のうち既定値である母音に対応する文字以外は持たず、既定値以外の母音音節は「既定値の母音文字 (V₀) + 母音記号」と表現する文字体系もある。この場合式5は更に式5 bのように書換えられることになる。ここでV₀は既定値の母音を表す母音文字である。実際タイ文字の体系には母音文字が存在せず、母音記号しか持たない。ミャンマー文字の体系においては母音文字も存在するが、基本文字表と考えられている34文字の文字表には既定値の母音文字しか含まれていない¹⁴。

<Vowel-Syllable>	::= V[D] V ₀ [M][D]	式 5 b
V ₀	: 既定値の母音を表す母音文字	

繰り返しになるが、この構文規則中で文字、記号の列が示すものはあくまでもそれらの「論理的構造」であり、空間的な位置関係、前後関係を示すもので

はない。またこの「論理的構造」は発話の過程において個々の音が生起する順序に対応していることから、音節の発音順序を示す「音韻的構造」と考えてもよい。

【図形表現】

（母音記号の配置）

母音記号は基底文字（子音文字）の上下左右に単独で或いは組み合わせられて配置される。配置される位置は母音記号の種類によって決まっており、これは既に表6に示した。本稿で取り上げた7つの言語をまとめて例示したものを表7に示す。特に注意を要するのは母音記号には基底文字の左側に配置されるものが存在することである。これは「先行母音」（leading vowel）と呼ばれ、符号化方式を考える場合に重要な役割を持つ。

表7 母音記号の配置位置（例示）

文字体系	先行母音 leading vowel	上付き母音 superscript vowel	下付き母音 subscript vowel	後続母音 following vowel	組み合わせ combination (surroundrant)
Devanagari	कि	कै	कू	की	
Khmer	កិ	ក្តិ	ក្និ	កឺ	ក្តិ ក្និ
Lao	ກິ	ກື	ກູ	ກີ	ກະ
Myanmar	ကိ	ကီ	ကု	ကု	ကို ကော
Sinhala	කි	කී	කු	කී	කො
Tamil	கி	கீ	கூ	கி	கோ
Thai	กิ	กี	กู	กิ	เก

子母音の結合による図形の合成に関して注意を要するのは以下の場合である。

第1に、母音記号には複数の表示形を持つものが存在することである。表6において同じセルに複数の記号があるのはこれを示す。この場合、結合する基底文字の種類に応じて母音記号の表示形が選択される。表示形の相違は配置位置の変化を伴うこともある。タミル文字の/u/, /ū/音を示す母音記号はそれぞれ2種類の表示形を持つが、一方は下付きの記号であり、他方は後置の記号である（表8参照）。異なった表示形が用いられる理由は様々であるが、（i）図形上の結合によって他の文字と誤読される惧れが生ずる、（ii）表示上空間

的混雑を引き起こす, (iii) 筆記を滑らかに行う上で無理を生ずる等の事情を回避するために行われることが多い。

第2に, 子母音の結合にあたって合字をつくることがある。タミル文字とデーヴァナーガリー文字の場合の合字例を表8及び表9に示した。特に変形が著しいのはタミル文字の場合であり, /ā/, /u/, /ū/及び/ai/を表す母音記号は多くの子音文字との間で合字をつくる。同表にはこれらの4種類の母音記号と全ての子音文字との結合による合成列を示した。表中, 合字と見なされる合成列を黒字で示した。

しかし第1と第2のケース区分は相対的なものである。例えばタミル文字 **கூ** の場合, これを子音文字 **க** と母音記号 **ூ** の合成により生成することも可能であるが, これに代えて一字の合字 **கூ** が用意されることもある。タミル文字のように手書き風な書体を持つ文字の場合, 子母音の境界は中断なく書かれることがあり, この場合レンダリングされた合成列は一筆で書かれたように見えることが要求される。フォント・メトリックスが適切に設計されればこの要求を満たす合成を行うことは可能であるが, 図形表記上の微妙な要求を満たすために合字が用意されることもある。

表8 Tamil文字の子母音結合における合字

v \ c	க	ங	ச	ஜ	ஞ	ட	ண	த
ஊ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
v \ c	ந	ன	ப	ம	ய	ர	ற	ல
ஊ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
v \ c	ள	ழ	வ	ஷ	ஸ	ஹ		
ஊ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ
கூ	கூ	கூ	சூ	ஜூ	ஞூ	டூ	ணூ	தூ

(注) 本表には合字を作らない子音文字の例(灰色で表示した部分)も含めて, 全ての子音文字を掲げた。

表9 Devanagari文字の子母音結合における合字例

子音文字	ह	र	र
母音記号	८	९	७
合字	हृ	रु	रु

(声調記号・補助記号の配置)

声調記号等の補助記号は基底文字の上下ないし後続位置に配置されるが、母音記号と組み合わせられる場合、一般的には最も「外側」に配置される。つまり上付きの母音記号と組み合わせられる上付きの声調記号の場合には母音記号の更に上側に、下付きの母音記号と組み合わせられる下付きの声調記号の場合には最下部に、という配置をとる。左側に配置される母音記号と左配置の声調記号が組み合わせられる場合には声調記号が最も左側に配置される（表10参照）。しかし空間的な混雑を引き起こす場合には、「最外側」という一般論に反して配置される場合もある。

表10 声調記号等の補助記号の配置位置（例示）

文字体系	先行 leading	上付き superscript	下付き subscript	後続 following	組み合わせ combination
Deavanagari	---	क̣	---	कः	---
Khmer	---	ក̣	ក̣	កៈ	កៈ
Lao	---	ກ̣ ກ̣	---	---	---
Myanmar	---	---	က̣	ကး	ကး
Sinhala	---	---	---	කො	---
Tamil	---	---	---	கொ	---
Thai	---	ก̣	---	---	---

3.3 多重子音の表現

【論理的構成】

インド系文字を用いて書かれる言語の中にはもともと多重子音を持たない言

語もあるが、これらの言語にあっても、サンスクリット語やパーリ語から語彙が輸入されるに従って多重子音の表現が必要となった。

結合音節文字では全ての子音文字は母音を伴う音節文字であるから、単に子音文字を重ねるだけでは「音節の連続」と解釈されてしまう。これを回避して多重子音を表現するために、インド系文字では多様な表現方法が発達してきた。現在の文字体系においては、多重子音を構成する各子音文字からの固有母音の削除を明示するため、

- ①対象となる子音文字に脱母音記号を付す
- ②脱母音された子音文字を表す別の記号を作る
- ③子音文字を垂直方向に重ねる（二階建て文字）
- ④複数の子音文字を組み合わせて合字を作る

といった表現が用いられている。表11に主なインド系文字体系においてどのような表現方法が用いられているのかを一覧にまとめた。

これらの表現方法を音節式として与えるために、構文規則における多重子音の音節式7を、子音文字、子音記号、脱母音記号から構成される構文規則として書き換えると次のようになる。なお、表記にあたっては一文字の略記号を用いた簡潔な表記を行っている。

<Cons-Conjunct>	::= C{XC}	: ①のケース	式7 a
	::= C{C} {C}C	: ②のケース	式7 b
	::= Cl	: ③, ④のケース	式7 c
C	:	子音文字 (consonants)	
C	:	子音記号 (consonant signs)	
Cl	:	合成された合字 (ligatures)	
X	:	脱母音記号 (devowelizer)	

しかしこの分類はあくまでも相対的なものであることを断っておかななくてはならない。例えばデーヴァナーガリにおける「合字」において、表11の例①のように合成前の字形を推測するのが困難なほどに変形を伴う合字もあれば、例②のように元の文字を想像できる程度の変形のものもある。この例の場合は「二階建て文字」と分類することもできよう。また例③における左側の子音文字 **𑖀** は合成にあたって文字の右半分が欠落した形となっているが、これは“Half letter”と呼ばれ、いわば水平方向に文字合成が行われる時の「部品」と

なっている。

クメール語においては全てを「二階建て文字」と分類したが、ほとんどの子音文字は二階建て文字の下付きとなる場合の特別の字形を持っており（表中④）、これらは「子音記号」と考えることもできる。

ミャンマー語においても、二階建て文字の合成にあたって、「合字」と呼ぶのが相応しい程の変形を伴う場合もある（表中⑤）。

表11 各文字体系における多重子音の構成法（例示）

文字体系	脱母音記号	子音記号	二階建て	合字
Devanagari	ट+ _゚ +ठ→टठ ड+ _゚ +ग→ङग	र+क→र्क	---	क+ष→क्ष ① ड+क→ङ्क ② क+र→कर ③
Khmer	---	---	ឃ→ឃ្ម ④ ខ	---
Lao	---	ນ+ຣ→ນຣ	---	---
Myanmar	န+ [်] →န့်	မ+ရ→မြ	က→ဏ က	က+ျ→ကျ ⑤ ခ
Sinhala	---	ක+ර→කර	---	---
Tamil	---	---	---	---
Thai	---	---	---	---

【図形表現】

(1) 脱母音記号

脱母音記号を用いて多重子音（ないし末子音）をあらわす方法はインド系文字を用いる言語体系で最も広く見られる方法である。

脱母音記号を用いて多重子音ないし末子音を表現する場合、脱母音記号は脱母音化する文字の上部又は下部に置かれる。

表12 インド系文字における脱母音記号

文字 名称	Devanagari Halant / Virama	Myanmar Athat	Sinhala Hal Kirima	Tamil Virama
字形	क्	ၵ်	ක්	க்

(2) 子音記号（半母音）

「半母音」（semivowels or medials）ともいわれる /ya/, /ra/, /la/, /wa/ の各子音は特に頻繁に結合子音を作るため、単独で音節を表す場合と異なる結合用の記号が存在する。日本語のカナにおける拗音を表す小文字の「ゃ、ゅ、ょ」は、その機能の点でも、またそれ自体単独で音節を表すことができない点でもこの子音記号に類似している。

インド系文字における子音記号を表13に掲げる。クメール文字の場合、全ての子音文字に「脚文字」（後述）と呼ばれる下付き記号が存在するが、この表には他の文字体系と共通する4つの音価に相応する記号のみを示した。ラオ文字では /y/, /l/ を表す二つの半母音だけが残し、シンハラ文字では /y/, /r/ を表す二つの半母音だけが残った。タミル文字及びタイ文字では半母音を表す子音記号は存在しない。

母音記号や補助記号と同様、子音記号も基底文字の上下左右或いは基底文字を包み込むように配置される。とりわけ表記上先行して配置される子音記号（表中①、②のケース。いずれも /r/ 音に相当する子音）は、先行母音と同様、符号化に際して注意を要する。

結合する子音が結合子音である場合、半母音を表す子音記号は結合する子音クラスターの全体を包み込む形となる。

表13 インド系文字における子音記号

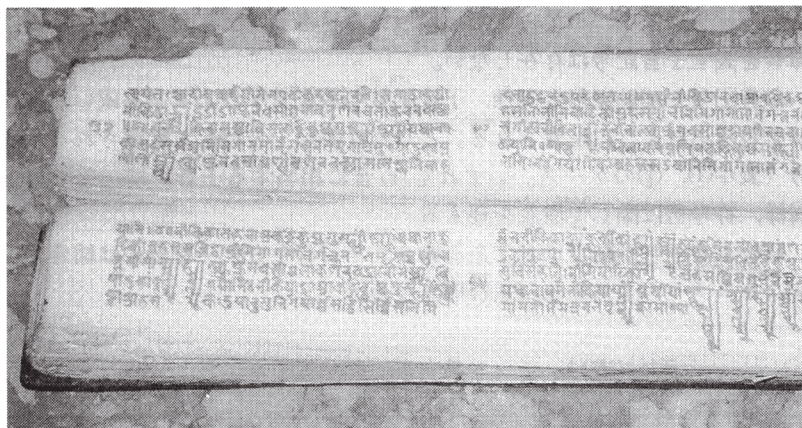
音価	"-y"	"-r"	"-r"	"-l"	"-w"
Devanagari	---	र्क्	र्	---	---
Khmer	រៀ	---	រ្រ ①	រ្ល	រ្វ
Laos	ນຸ	---	---	ນ	---
Myanmar	ျ	---	ြ ②	့	ွ
Sinhala	ක්‍ය	---	ක්‍ර	---	---

（注）"-r"は、基底文字となる子音に対して /r/ が発音上先行する場合を表し、"-r"は後置される場合を示す。

(3) 二階建て文字

子音文字を垂直方向に重ねて書くことにより結合子音であることを示す方法は、デーヴァナーガリー文字、チベット文字、カンボジア文字、ミャンマー文字など多くのインド系文字に見られる。子音文字が水平に連ねて書かれるときにはそれぞれが独立の音節を表すものとして読まれ、垂直に連ねて書かれるときには一つの音節として読まれる。

こうした「二階建て」記法はサンスクリット語にその起源を持つものといわれている。筆者はネパールの古文書館に保存されたデーヴァナーガリー文字で書かれた古文書（約1200年前に書かれたもの）に七重の多重子音が記されているのをみたことがある（写真参照）。この場合、文字列は多重子音のところだけ下に長く垂れ下がった形で書かれる。現代のデーヴァナーガリー文字に見られる ष, ॠ, ॡ, ॢ, ॣ といった合字は合成後の文字の高さが通常の子音文字と同じ高さに圧縮されているが、古代に行われた「垂直重ね書き」の痕跡とみることもできよう。



写真：デーヴァナーガリー文字で書かれた経典
（ネパール、ケイサー図書館所蔵）

二階建て文字における下付きの子音文字を子音記号と見なせば、前項(2)で述べた子音記号による多重子音の表現と同形である。但し「二階建て」としての制約から三重以上の多重子音を表現することができないことは当然である。

「二階建て」の場合、字形が変化しなくても結合文字であることは明らかであるが、空間が限られていることから、①総じて字形が小さくなる、②（字形が縦長の場合などにはそのままでは重ねることができないため）上付きとなる文字の脚の部分小さくする、或いは③下付きとなる文字を変形したり回転させたりする等の字形変化が生ずる。この変位、変形は時に原形を止めないほどであり、この場合には次項で述べる合字と考えるのが自然である。

二階建て文字の実例1：カンボジアの胴文字・脚文字

カンボジア文字には33文字の子音文字がある。全ての子音文字に「胴文字」と「脚文字」（choeng pyany-cha'-na'）という二つの形があり、二重子音を作る時には先に発音される子音を胴文字で書き、後に続いて発音される子音はその下に脚文字で書く。脚文字は胴文字の簡略形である場合が多いが中には胴文字の形からは類推できないほど変形してるものもある。この二階建ての子音の上下左右に母音記号や声調記号が付加されて一つのまとまりを作る。

表14 カンボジアの胴文字と脚文字

音価	ka	kha	ko	kho	ngo	cha	chha	cho	chho	nyo	da	dha	do	dho	na	ta	tha
胴文字	ក	ខ	គ	ឃ	ង	ច	ឆ	ជ	ឈ	ញ	ដ	ប	ព	ផ	ណ	ត	ថ
脚文字	ក	ខ	គ	ឃ	ង	ច	ឆ	ជ	ឈ	ញ	ដ	ប	ព	ផ	ណ	ត	ថ
音価	to	tho	no	pa /ba	pha	po	pho	mo	yo	ro	lo	vwo	sa	ha	la	a	
胴文字	ទ	ធ	ន	ប	ផ	ព	ភ	ម	យ	រ	ល	វ	ស	ហ	ឡ	អ	
脚文字	ទ	ធ	ន	ប	ផ	ព	ភ	ម	យ	រ	ល	វ	ស	ហ	ឡ	អ	

（注）脚文字の“da”と“ta”，“la”と“lo”は同形であるため、脚文字は31種類しかない。

なお/r/音を表す脚文字はそれが付加される基底文字よりも左側に書かれることに注意する必要がある。また希に三重子音を作ることがある。次の第2例

と第3例は/r/音を含む3つの子音文字からなる多重子音の例である。/r/音が左側に書かれるため三階建てとはならないが、/r/音脚文字の脚部分は通常よりも下に伸びている。

例 カンボジア文字の多重子音の例（黒字部分が多重子音）

例	ស្រា	កន្ទ្រី	សូន្យក្រោម
音価	sra	ka-ntrai	so-ngkreām

(4) 多重子音の表現：合字 (Ligature)

多重子音が合字 (Ligature) として表現されることがある。(1)項で述べたようにデーヴァナーガリー文字においては多くの多重子音が合字で表現される。その際”Half letter form”と呼ばれる、文字の右半分を取り去った形の図形が多用される。表15にこれを掲げる。Half letter formを作る子音文字のほとんどは右側に「r」を持つ文字であり、この部分は既定値母音/a/を表しているから、これを取り去ることにより脱母音された状態を示す、と解釈することもできよう。子音文字の組み合わせによってできた合字自身が Half Letters Formとなることもある。この事例を表15bに掲げた。

表15a Devanagari子音文字のHalf Letters Form

IPA	ka	kha	ga	gha	nga	cha	chha	ja	jha	nya	da	dha	da	dha	na
Full form	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	ठ	ड	ढ	ण
Half form	क्	ख्	ग्	घ्		च्	छ्	ज्	झ्	ञ्					ण्
IPA	ta	tha	ta	tha	na	pa	pha	ba	bha	ma					
Full form	त	थ	द	ध	न	प	फ	ब	भ	म					
Half form	त्	थ्		ध्	न्	प्	फ्	ब्	भ्	म्					
IPA	ya	ra	la	lla	va	sha	ssa	sa	ha						
Full form	य	र	ल	ळ	व	श	ष	स	ह						
Half form	य्		ल्	ळ्	व्	श्	ष्	स्	ह्						

表15b Devanagari子音結合文字のHalf Letters Form

IPA	ka+ssa	cha+na	ta+ta
Full form	क+ष→क्ष	ज+ञ→ज्ञ	त+त→त्त
Half form	क्ष्	ज्ञ्	त्त

4 結合音節文字の符号化方式

4.1 符号化方式の3類型

以上により結合音節文字における符号化方式を論ずるための準備が整った。結合音節文字の符号化方式については、符号化の対象となる基本単位をどのように捉えるかという観点から3つの大きなカテゴリーに分類することが可能である。第一は音節あるいは文字列を構成する個々の図形的要素を基本単位としてコード化する方式であり、これをGraphic Codingと呼ぶ。第二は音節を構成する個々の音韻的要素を基本単位としてコード化する方式であり、これをPhonetic Codingと呼ぶ。第三は音節を図形的要素或いは音韻的要素に分解せずにそのまま音節を単位としてコード化する方式であり、これをSyllabic Codingと呼ぶ。以下、各方式について逐次解説していく。

(1) 図形要素を単位とする符号化 (Graphic Coding)

この方式による場合、個々の図形要素を基本単位として符号化を行う。形状或いは表示位置の異なる図形には異なる符号化表現を与えるというのがこの方式の基本原則である。従ってこれを結合音節文字に当てはめれば、次のような符号化が行われることとなる。

- ①母音文字と母音記号は同じ音価を示すものであっても、異なる図形表現を持つ限り異なる符号化表現を与える
- ②同様にして子音文字と子音記号も異なる図形表現を持つ限り異なる符号化表現を与える
- ③脱母音記号のような制御記号的な機能を有する記号についても符号化表現を与える

この符号化方式は、二次元的に配置されている図形要素をどのような順序で符号列上に配列するかという観点から更に幾つかの方式に分かれる。

GCGO : Graphic Coding Graphic Order

第1の方式は図形要素を、文字列の進行方向を基準とした順序に従って符号列に並べていく方法である。図形が上下に重なっている場合には何らかの規則を定めて一次元的な符号列に変換する。先行母音や先行子音が存在するために、符号列上に現れる符号化表現の順序は発音順序と対応しないことがある。

GCWO : Graphic Coding Writing Order

第2の方法は、図形要素をその筆記順序に従って配列していく方法である。

筆記という動作は発話動作と同様に時間の経過とともに進行する一次元的な過程であるから、こうした配列は原理的には可能である。しかし筆記順序が教育課程において普遍的に教育され共通の知識となっていない場合には、この方法は普遍性のある符号化方式とはならない。

GCPO : Graphic Coding Pronunciation Order

第3の方法は、図形要素を符号化の単位とするが、その符号列への配列にあたっては発音上の順序に従う方法である。

(2) 音価を単位とする符号化 (Phonetic Coding)

第二の方式は、表記上の個々の図形要素が持つ音価を基本単位として符号化を行う方式である。同一の音価を有する文字・記号には同一の符号化表現を与えるというのがこの方式の基本原則である。具体的には

- ①子音文字に含まれる既定値母音にも明示的に符号化表現を与える
- ②母音文字と母音記号は区別せず、音価の同じものは同一の符号化表現を与える
- ③脱母音記号は符号化を行わない
- ④ɸ音記号も子音文字と区別しない
- ⑤複数の図形記号によって一つの母音が表されている場合 (surroundrant vowels) であっても、一つの音価を表すものであるから一つの符号化表現で表す

となる。音節文字としての特徴を捨てて、表音文字的なコード化を行うものであるともいえる。一般に図形要素を単位とする符号化方式の場合に比べて、必要とする符号の種類は少なくなる。しかしこの方式をとる場合、個々の符号化表現は表記上の文字・記号と対応しなくなるために、入力時点及び出力時点で必ず変換を必要とする。インドNCST (National Center for Software Technology) のS.P.Maduraは、Graphical CodingをScript Level Codingと呼び、この方法をLanguage Level Codingと呼んでいる。

なおここで注意を要するのは、この方式は発音そのものの符号化とは異なるということである。例えば綴り字の上で発音されない文字や音便規則に従って実際の発音が変化する場合、発音に忠実に従った符号化を行えば、前者は無視され、後者は実際の発音に対応した音価の符号化表現として表される。こうした符号化も考えられるが、これはもはや文字の符号化方式とは呼べない。

(3) 音節を単位とする符号化 (Syllabic Coding)

以上の二つの方式と異なり、音節をそれ以下の基本単位に分解せず、音節そのものを基本単位として符号化を行うという方法もありうる。当然のことながら、その場合には符号化すべき基本単位の数は飛躍的に大きくなるが、他方で結合文字の扱いに関連する複雑さは一切無くなる。音節文字をそのまま符号化する方式といえる。

その際、音節の総数が幾つになるかは言語によって異なるが、一般的には数千のオーダーとなり、必要なコード・ポジションの数は1バイト符号表に収容可能な大きさをはるかに超え、2バイト符号表とせざるを得なくなる。声調言語の場合であって、しかも声調の異なる音節を異なるセルに割り当てることとする場合には、必要なセル数は更に大きくなる。

4.2 GCGO: Graphic Coding Graphic Order

前節の分類法に基づいて、国家的レベルで文字符号化規格が定められている幾つかの事例を分類していこう。

GCGOによる場合、図形要素はその形状及び表示位置によって区別されるから、母音記号や補助記号のように様々な位置に表示されるものはその表示位置に従ってカテゴリー分けが行われる。例えば母音記号に先行、上付き、下付き、後置の4種類があり、補助記号が上付きであるとする。そして同一カラム内の符号化順序として、下付き→上付き→最上付きという順序を定めたとする。ここで先行母音を V_L 、下付き母音記号 V_s を、上付き母音記号 V_A を、後置母音記号を V_T 、上付きの補助記号を D_T とすると、音節の符号化表現を示す式は次式のように表される。

$$\langle \text{Syllable} \rangle ::= [V_L]C[V_s][V_A][D_T][V_T] \quad \text{式 8 b}$$

V_L	:	先行母音記号
V_s	:	下付き母音記号
V_A	:	上付き母音記号
D_T	:	上付き長音記号
V_T	:	後置母音記号

一般化して考えれば、GCGOは音節ではなく進行方向に垂直な「カラム」を単位として符号化を行う方式である。従ってある記号が母音記号であるか、補

助記号か、子音記号であるかの区別は二次的な区別であり、それがどの位置に配置される記号であるかで分類される。こうした特徴を明示的に表現して符号化方式を記述すれば次式 8 c のようになる。

この方式はColumn-wise codingと呼んでもよい。

<Column> ::= B[S][A][T] 式 8 c

B	:	基底文字
S	:	下付き記号
A	:	上付き記号
T	:	最上付き記号

実例 1 タイの文字コード規格：TIS 620-2533

タイでは1986年（タイ暦2529年）にタイ文字の符号化に関する国家規格 **TIS 620-2529: Standard for Thai Character Codes for Computers** が制定され、4年後の1990年（タイ暦2533年）に改訂されてTIS 620-2533となった。**TIS 620-2533**は子音文字44種類、母音記号32種類、タイ数字10種類、調音記号4種類、その他記号4種類の合計94種類の文字・記号を収録している。

この規格はGCGO方式に分類されるものであり、図形要素のコード列上での順序については、文字列の進行方向に即して登場する順序に従っている。先行母音はそのまま基底文字に先行してコード列上に現れる。垂直方向に記号が重なる場合には、上から4段階のレベル分けを行い、基底文字（第3レベル）→下付き記号（第4レベル）→上付き記号（第2レベル）→上付き記号（第1レベル）という順序でコード列上に並べられる。TIS 620自身はこの順序に関して特段の定めを置いていないが、符号化規格の実装ガイドラインというべき文書**WTT 2.0**において、このような規定が置かれている。また同文書には、コード列上のこうした順序を確保するために、入力FEPにおいてここに述べたような順序が確保されるようなチェックを行うことを推奨している。

4.3 GCPO: Graphic Coding Pronunciation Order

一方GCPO方式による場合、符号化の基本単位は図形要素であるが、その順序は発音順序に従う。従って開音節の場合の符号化表現は音節の論理構造式と同一である。

<Vowel-Syllable> ::= V[D]

<Cons-Vowel-Syllable> ::= C[V][D]

多重子音の符号化表現についても、「表示形或いは表示位置の異なる図形要素には異なる符号化表現を与える」というGCPOの基本原則に従うならば、下付きの子音記号や半母音のための子音記号には子音文字と別の符号化表現を与えることになる。従って上記の式を一般化した表現は次式のようになる。

<Vowel-Syllable> ::= V[D]

<Cons-Vowel-Syllable> ::= C{C}[V][D]

実例2 シンハラ文字に関するスリランカ国家規格：SLS 1134:1996

この符号化方式に従う実例がスリランカの国家規格SLS 1134である。この規格の定める符号化方式はGraphic Codingに従っている。母音文字と母音記号は区別され、別の符号化表現が与えられている。またシンハラ語には多重子音があるが、多重子音を作るために用いられる子音記号は子音文字とは別の符号が与えられている。








またこの規格においては符号列上での図形要素の配列順序に関して、入力順序や表記順序と関係なく、論理的・音韻的構造を優先させていることである。表16にその様子を示す。母音記号（）は子音に先行して現れる母音記号であり、入力時点でも筆記順序に従って先頭に書かれるにもかかわらず、内部符号列上では子音文字（）、子音記号（）、母音記号（）、母音記号（）、脱母音記号（）の順序に従って配列されている。入出力の際には置換処理（permutations）が要求される。

図16 SLS 1134における先行母音（）の置換処理

入力順序	内部コード列					出力文字
						
 						 
 						 

（出典）S.T.Nandasara & V.K.Samaranayake, Current Development of Sinhala/Tamil/English Trilingual Processing in Sri Lanka, The Second MLIT Proceedings, Nov. 7-8, 1997, Tokyo

4.4 GVPC: Graphic Vowel-coding Phonetic Consonant-coding

GC方式の変形として、母音の符号化はGC方式に従い、子音の符号化はPC方式に従うという事例もある。これはインドの国家規格IS 13194である。

実例3 インドの公用文字9種に関するインド国家規格：IS 13194:1991

インド文字の符号化構文規則については、IS 13194:1991が明示的な形でBNF形式の表現を与えている¹⁵。以下にこれを原典のまま記す。

Word	::= {Syllable}[Cons-Syllable]	式9 a
Syllable	::= Vowel-Syllable Cons-Vowel-Syllable	式9 b
Vowel-Syllable	::= V[D]	式9 c
Cons-Vowel-Syllable	::= [Cons-Syllable]Full-Cons[M][D]	式9 d
Cons-Syllable	::= [Pure-Cons][Pure-Cons]Pure-Cons	式9 e
Pure-Cons	::= Full-Cons H	式9 f
Full-Cons	::= C[N]	式9 g

式9 aは、子音で終わる単語（閉音節語）の存在を示したものである¹⁶。

式9 bは、音節が母音のみの音節か或いは子音を含む音節のいずれかであることを示す。

式9 cは、母音音節に含まれる母音は一つであり、これに声調記号が付加される場合がありうることを示す。

式9 dは、子音を含む音節の基本構成を示している。それは必ず一つのFull-Consを含み、音節主音となる母音はこのFull-Consに含まれる母音によって与えられる。しかし母音変更記号Mが付加されている場合には、既定値の母音がMの与える音価に変わり、これが音節主音となる。声調記号が付加されることは第3式と同様である。

式9 eは、子音音節における多重子音の構成を示す。式9dに含まれるFull-Consを含めれば、多重子音は最大で4重までであることが示されている。

式9 fは、多重子音を構成する場合に、子音を脱母音記号Hによって挟むことによって表現するものであることを示す。

式9 gは、Full-Consが基本的には子音文字で与えられるが、場合によってはNukta記号Nが付加されることを示す。

この方式を実例2で示したGCPO方式と区別する最大の特徴は、子音文字と

子音記号とが区別されていなことである。このため式9 f に示されるように脱母音記号を明示的に符号化し、これによって脱母音された子音を表現している。既に前項で述べたように多重子音の表現方法としては脱母音記号によらずに合字や子音記号による場合も含むが、この方式はこれら全ての場合を明示的な脱母音記号の挿入によって表現している。

4.5 PC: Phonetic Coding

Phonetic Codingに従う場合、符号化の構文式は次式のように表される。単独の子音文字で表される音節も含めて、音節は必ず母音を含むから母音が省略されることはない。式10と式11を併せて書けば式12となり、むしろ子音が省略されることのある表現となる。

ここで符号化される単位は文字ではなく音価となるから、表記にあたってGraphic Codingのケースと区別するために小文字を用いる。子音を表す符号はcであり、子音文字、子音記号ともに、同じ音価であれば同じ符号を用いて表す。同様に母音を表す符号はvであり、母音文字も母音記号も同じ音価であれば同じ符号を用いて表す。また多重子音は子音を単純に重ねることにより表現されることになる。

閉音節を含む場合、この符号化方式はもう少し複雑となるが、ここではそのことに立ち入らない。

<Vowel-Syllable>	::= v[d]	式10
<Cons-Vowel-Syllable>	::= {c}v[d]	式11
<Syllable>	::= [{c}]v[d]	式12

但し

$$c = C = \underline{C} \quad (\text{子音の音価が同じ})$$

$$v = V = \underline{V} \quad (\text{母音の音価が同じ})$$

実例4 インドの旧国家規格: ISSCII-83

Phonetic Coding方式に近い実例は現行インド国家規格IS 13194に先立って1983年に定められた文字符号化規格ISSCII-83 **Indian Script Standard Code for Information Interchange**である¹⁷。符号化構文規則は次式のとおりである。なお、本稿における記号の用法と合わせるために若干の書き換えを行っている。

Syllable	::= Vowel-Syllable Cons-Syllable SYM	式13
Vowel-Syllable	::= LNK v[S̃]	式14
Cons-Syllable	::= [c LNK][c LNK][c LNK] c [v][S][S]	式15

SYM	:	非結合記号
LNK	:	リンク記号
v	:	母音を表す記号
c	:	子音を表す記号
S	:	補助記号

ここで特徴的な点を挙げて補足しておく。

第1に、母音は単独で用いられる母音文字と、結合文字である母音記号とが区別されず、両方とも記号「v」で表されている。出力時点では、母音が音節頭に置かれる場合には母音文字に、子音文字の後に置かれる場合には母音記号に変換する処理が必要である。符号列からの変換規則としては、LNK記号の後の場合は母音文字、子音「c」の後の場合は母音記号、という規則となる。このことにより母音の表現に必要なコード・ポジションの数は約半分となっている。

第2に、多重子音は子音を表す記号の間にリンク記号を挿むことによって表される。子音とリンク記号からなる符号列は出力時点では合字或いは明示的な脱母音記号を含む図形記号列に変換されなければならない。

上記において「S」は3種類の補助記号（Anuswar, Chandrabindu, Visarg）、脱母音記号（Halant）、ヌクタ（Nukta）及びレフ（Reph）の6種類の記号を含む。これらの記号は組み合わせられて用いられることもあるために第3式右辺末尾には[S] [S]と二重に現れる。しかしこれらの記号が音節結合規則上果たしている役割は様々であり、これらが単純に後置されているのは余り論理的とは言えない。

なおこの規格が純粋なPhonetic Coding方式と異なるのは、式15において母音の省略がありうるとしている点である。これは部分的にGraphic Codingの考え方をとっているためである。しかしこのために音節頭の母音や多重子音を識別するための何らかの追加情報が必要となり、リンク記号「LNK」を用いた識別を行うこととしている。多重子音については図形表現の上でも脱母音記号を書く習慣があるためにLNKの入力に然程の抵抗はないかもし得ないが、

頭母音の入力に先立つLNK記号の入力はおそらく違和感のあるものであったのではなかろうか。

4.6 SC: Syllabic Coding

音節を単位とする符号化方式は音節内の構造に立ち入らないから、符号化方式は単純である。現在のところ音節文字に関する文字符号化規格の中で完全にこの方式に従っているのは韓国のハングル文字符号化規格である。日本語の文字符号化規格も促音と撥音を例外と見なせばこの方式といえる。

既に述べたように、この方式による場合、符号化に必要なコード表の大きさはかなり大きくなる。一例としてミャンマー語の綴字法辞典¹⁸に列挙された音節表を分析すると、あらゆる子母音、補助記号の可能な組み合わせから計算される音節の総数は3,500余りとなる（表17）。

表17 ミャンマー語における音節数

	開音節	閉音節	合計
単子音(注)	C[M][D] 396	C[M][D]CH 792	1,188
二重子音	CC[M][D] 580	CC[M][D]CH 1,148	1,728
三重子音	CCC[M][D] 232	CCC[M][D]CH 397	629
合計	1,208	2,337	3,545

(注) 「a」音の母音文字も子音文字と見なされているために母音で始まる音節もこの中に含まれている。ちなみに、音節頭が母音で始まるものの数は合計79種類であった。

5 まとめ

本稿では、インド系文字に関してこれまでに制定されてきた国家規格を対象としてその符号化方式の分類を試み、

①符号化の単位のレベル（図形要素、音価、音節）

②符号化データを配列する順序（図形上の順序、発音順）

の二つの基準に従って5種類の類型に分類した。またそれぞれの符号化方式を、音節の論理的構造式によって表現した。結合音節文字の場合、符号表が与えられただけでは符号化方式を十分に理解することができないが、本稿のような形式的表現を用いて符号化方式を記述することにより、十分な言語知識を持たない情報処理の専門家にとっても理解が容易になることを示し得たのではないかと考える。

本研究の最終的な目的は、文字符号化方式を情報処理の観点から分類類型別に評価し、その長所・短所を明らかにして望ましい符号化方式についてのガイドラインを作成することであるが、以上の分類によって得られる知見によっても幾つかの評価を得ることができる。詳しくは別稿をもって論ずるが、今後の検討課題も含めて簡単にこれをまとめれば表18となる。

表18 符号化方式の特徴

方式	Graphic Coding			Phonetic Coding	Syllabic Coding
	GCGO	GCPO	GVPC	PC	SC
実例	TIS620	SLS1134	ISCI	ISSCI83	KSC5601
符号化文字集合	{C}, {C̣} {V}, {Ṿ} {D}, X {Sn, Sp, Sm}	{C}, {C̣} {V}, {Ṿ} {D}, X {Sn, Sp, Sm}	{C} {V}, {Ṿ} {D}, X {Sn, Sp, Sm}	{c} {v} {D}, X {Sn, Sp, Sm}	{S} (音節) {Sn, Sp, Sm}
符号表の大きさ	通常シングルバイトでの符号化が可能な大きさ				音節数は通常数千に及ぶためマルチバイト化が不可欠
入力時のWYSWYG性	進行方向に沿った順序のままで入力	先行母音等については符号化文字列を生成する時に置換が必要であり、WYSWYG性が低い		音素への変換を行いながら入力する必要	WYSWYG性はあるが、入力方法に工夫が必要
出力時のレンダリング負担	小さい	小さい	多重子音についてレンダリング負担あり	レンダリング負担が大きい	なし
テキスト編集時のWYSWYG性	WYSWYG性が高い	置換が行われているためにWYSWYG性が低い	多重子音に関してはWYSWYG性が低い	WYSWYG性が低い	WYSWYG性が高い
メモリ効率	今後の検討課題				
辞書順への直接ソートの可能性	今後の検討課題				

ここで、「入力時のWYSWYG性(What You See is What You Get)」及び「テキスト編集時のWYSWYG性」とは、文字列の当初入力操作時及び加除修正時における「自然さ」を表す評価尺度であり、「出力時のレンダリング負担」とは、内部符号列から出力文字をレンダリング(生成)する際の変換ソフトウェアへの負担を表す評価尺度である。文字符号化方式は情報処理システム内部における表現を規定するものであるが、入力時の「自然さ」と出力時のソフトウェアへの「負担」は、内部表現方法の選択がシステムの利用者及びシステムそれ自体にどのような負担を課すかを見極める尺度である。表18に示したとおり各方式はそれぞれに一長一短があり、ここまでの検討結果だけで符号化方式の優劣を論ずることはできない。「自然さ」や「負担」といった尺度をどのようにしてより客観化していくのか、という課題も残されている。また、本稿の分析では何らふれていない「メモリー効率」や「辞書順序への直接ソートの可能性」といった評価視点も取り入れる必要がある。

「メモリー効率」とは、ある長さの文字列を符号化した場合に、情報処理システム内部でそれがどの程度のビット数で表現されるかを表す指標であり、「音節あたり平均ビット数」等の指標で計測できるものと考えられる。但しこの計測にあたっては本稿で与えた音節の論理構造式が重要な役割を果たすものと予想している。

「辞書順序への直接ソートの可能性」とは、符号化文字列に対して直接ソート(バイナリー・ソート)を行った場合に、ソート結果が通常の辞書式順序に整列されるか否かを意味するものであり、符号化方式設計上の重要な基準の一つである。この点に関する検討は、まず各言語における辞書式順序の定式化から始めなくてはならず、今後の研究課題であるが、ここでも音節の論理構造式が重要な手掛かりになるものと考えている。

なお本稿においては、複数の表示形をもつ文字の扱い(論理構造式の上では全く区別されないが、表示上の都合により図形表現のことなる文字・記号)や"sarroundrant"な母音記号(表7参照)の取扱に関する議論を省略した。この点もいずれ別稿で論ずることとしたい。

参考文献

【国家規格】

- ☐ [Bangladesh] Bangladesh Standards and Testing Institution, BDS 1520: 1995, *Bangla Coded Character Set*
- ☐ [Cambodia] Cambodian Information Technology Association (CITA) meeting memos, URL <http://www.bigpond.com.kh/users/moritta/KhmTable.html>
- ☐ [China] *Chinese Character Set for Information Interchange*, GB2321: 1980
- ☐ [India] Bureau of Indian Standard (BIS), *Indian Script Code for Information Interchange*, ISCII: 1991.
- ☐ [Japan] Japan Industrial Standard Council (JISC), JIS X0201, JIS X0208, *Standard Character Code for Information Interchange*. (in Japanese)
- ☐ [DPRK] Committee for Standardization of DPRK, *Standard Korean Graphic Character Set for Information Interchange*, KPS 9566: 1997
- ☐ [ROK] Korean Standard KS C 5601-1987
- ☐ [Nepal] Nepali Font Standardization Committee (1998), *Nepali Font Standards White Paper Ver.2*
- ☐ [Sri Lanka] Sri Lankan Standard, SLS 1134: 1996
- ☐ [Thailand] Thai Industrial Standard Institute (TISI), TIS 620-2533(1986): *Standard for Thai Character Codes for Computers*. (in Thai)
- ☐ [Tamil] Announcements of Honorable Chief Minister of Tamilnadu during the International Conference "TAMILNET 99" held in Chennai, India during Feb.7 and 8, 1999.
- ☐ [Vietnam] *Vietnamese Standard Character Code for Information Interchange*, TCVN 5712-1993 (VSCII)

【国際規格】

- ☐ [International] ISO/IEC (1993), ISO/IEC 10646-1, *Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane*.
- ☐ [International] Unicode Consortium (1996), *The Unicode Standard, Version 2.0*. Addison-Wesley.

【その他の文献】

- ☐ Breton, Roland (1997), *Atlas of the Languages and Ethnic Communities of South Asia*, Sage Publications, New Delhi.
- ☐ Center of International Cooperation for Computerization: CICC (1994), *Cumulated Proceedings of the Asian Forum for Standardization of Information Technology 1987-1993*, URL <http://www.cicc.or.jp>
- ☐ CICC (1999), *Proceedings of the 3rd International Symposium on Multilingual Information Technology (MLIT-3)*. URL <http://www.cicc.or.jp>
- ☐ Comrie, Bernard (1990), *The Major Languages of East and South-east Asia*, Routledge, London.

- ☐ Coulmas, Florian (1996), The Blackwell Encyclopedia of Writing Systems, Blackwell Publishers, London.
- ☐ Mackenzie, C. E. (1980), Coded Character Sets, History and Development, Addison-Wesley.
- ☐ Nakanishi Akira(1980), Writing Systems of the World: Alphabets, Syllabaries, Pictograms, Charles E. Tuttle Co.
- ☐ Network Wizards(1999), Domain Name Survey, July 1999, URL <http://www.nw.com>
- ☐ 言語学大辞典, 三省堂, 1989年

註

- 1 長岡技術科学大学 計画・経営系
- 2 ここでの用語定義は三省堂言語学大辞典による。但し「結合音節文字」という表現は言語学の分野で一般に承認された術語ではない。
- 3 1999年現在, この国際共同プロジェクトには17カ国(カンボジア, 中国, インド, インドネシア, 日本, 韓国, ラオス, マレーシア, モンゴル, ミャンマー, ネパール, パキスタン, フィリピン, シンガポール, スリランカ, タイ, ベトナム)が参加している。
- 4 国際的な情報技術関係企業のコンソーシアムであるUNICODEが開発を進めている2オクテット或いは4オクテットの多言語文字符号化規格。1991年にその第1版が出版された。しかしこの版が収録しているアジアの文字は日本語/中国語/韓国語で共通に使用する漢字(統合漢字と呼ばれる), ハングル, カナ, インドの公用語で使用する9種類の文字, タイ文字, ラオ文字のみである。
- 5 国際標準化機構であるISO/IEC JTC1で開発が進められているマルチオクテットの多言語文字符号化規格。1993年に第1版が出版された。第1版にはUNICODEと同様の文字が収録されている。
- 6 各国の専門家による検討成果発表の場として, 1997年以降毎年開催されている“International Symposium on Multi-lingual Information Processing”(MLITと略称される)がある。各回の論文集については参考文献欄参照。
- 7 表1に掲げた範囲では, ベトナム, マレーシア, インドネシア, フィリピン, ブルネイが例外となる。しかしこれらの地域でローマ字化が浸透する以前は, 広くインド系文字が使用されていたと考えられている。
- 8 ISO/IEC 10646では「文字」とは「データの構成, 制御又は表現に用いる要素の集合の構成単位」(本文4.6項)と定義されており, 言語学或いは通常の文字観念よりもはるかに広い概念を包含する。また, 結合文字(combining character)とは「この規格群で規定する符号化文字集合の識別された部分集合の構成単位であって, 先行する非結合文字(以下, 基底文字という。)と組み合わせることを意図したもの, 又は基底文字の後に結合文字の列が続いた形のもの」と組み合わせることを意図したもの」(本文4.11項)と定義され, 合成列(composite sequence)とは「基底文字とそれに続く一つ以上の結合文字とからなる図形文字の列」(同4.13項)と定義される。合成列は文字とはみなさない。
- 9 英語でもB, C, D, G, P, T, Zは同じ母音を付けて発音するし, F, L, M, N, S, Xについては同じ頭母音を付けて発音する。しかし実際の音節中に現れた場合, このような母音音価を持っていると見なされているわけではない。
- 10 声調とは「語の意味を区別するのに役立つような声の高さの段階, またはその変化のうちで, 個々の音節について定まっているものをいう。声調を持つ言語は声調言語と呼ばれる」(三省堂言語学大辞典)。

- 11 音節の中心となる音素を音節主音と呼ぶ。音節主音は基本的には母音であるが、例外的には子音が音節主音となることもある。(三省堂「言語学大辞典」の「音節構造」の項)
- 12 日本語のカナも、濁音、半濁音、拗音、撥音等の表記において結合音節文字としての性質を持つが、これらの表記の必要は歴史的に発生したものであり、古来の日本語は完全に音節を単位とする表音文字であった。これは「カナの原型である漢字が元来単音節の表語文字であり、漢字の意味を捨象してその音だけを利用したところから発したカナは、音声的に音節を単位とするようになったのである」(前掲「言語学大辞典」の「音節文字」の項)と説明される。
- 13 音声としての言語が視覚的な形象としての文字に写されたことに伴う大きな変化がこの問題である。「視覚的印象の特徴が、時間的に先後して一次元的に展開するところにあるのに対し、視覚の方は、空間的に三次元的あるいは二次元的に構成される特徴がある。前者は連続して流れていくが、後者は部分に区分されて意識される。このようなはなはだしい相違を乗り越える言語の文字化は、必然的にいろいろな問題を含んでいる。」(前掲「言語学大辞典」の「文字」の項)
- 14 ミャンマー語の子音表のことを「アカラ」(akkhara)ないし「カージー・カーグエ」(Ka.ji: hka.gwei)と呼ぶ。「アカラ」はパーリ語で「不滅のもの」の意味を有し、「一旦口からでた言葉は不滅である」といった観念が根本にあるといわれる。この子音表には母音既定値である/a/音の母音文字だけが含まれており、他の母音は含まれていない。なお、「カージー・カーグエ」は文字表の先頭二文字の名前を続けたものであり、いわば「アルファ」と「ベータ」に相当する。クメール文字、ラオ文字、タイ文字においても、ミャンマー文字と同様、子音表の最後に母音既定値である/a/音の母音文字だけが含まれている。
- 15 IS 13194:1991の8.1項を参照。
- 16 語中、語頭の音節には閉音節が現れないことを示したものも言える。語頭、語中の音節の末子音は次音節の頭子音と結合した多重子音として扱われるため、閉音節となるのは語末の音節だけとなる。
- 17 Indian Script Standard Code for Information Interchangeの略。現行の標準はISCIIと略され、Sが一つ少ない。紛らわしく注意を要する。
- 18 かつて寺子屋で教えられていたミャンマー語教科書の復刻版(မြန်မာသင်ပုန်းကြီး)によった。